

University of Groningen

Component analysis of multisubject multivariate longitudinal data

Timmerman, Marieke Engelen

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version

Publisher's PDF, also known as Version of record

Publication date:

2001

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Timmerman, M. E. (2001). *Component analysis of multisubject multivariate longitudinal data*. s.n.

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Summary in Dutch (Samenvatting)

Het kenmerk van longitudinaal onderzoek is dat gegevens verzameld worden op verschillende tijdstippen bij hetzelfde individu of object. Met het verzamelen van longitudinale gegevens beoogt men inzicht te krijgen in het verloop van een bepaald proces, dat in het algemeen aan verandering onderhevig zal zijn. Om inzicht te krijgen in het verloop van een proces is het nodig om dit proces te meten. In de sociale wetenschappen is het vrijwel nooit doenlijk om een proces continu te meten, en zal een beperkt aantal meetmomenten gekozen worden. Belangrijke overwegingen bij de keuze van die meetmomenten zijn de mate van precisie waarmee een proces gevolgd moet worden om de onderzoeksvragen te kunnen beantwoorden, de kennis die bij voorbaat aanwezig is over de vorm van het te volgen proces, en de grootte van de meetfout.

Volgt men de processen van meerdere individuen, dan kan men niet alleen de individuele ontwikkeling van de afzonderlijke individuen onderzoeken, maar ook de interindividuele verschillen in ontwikkeling. Hierbij is het zinvol om twee typen onderzoeksdesigns te onderscheiden. Ten eerste, het design waarbij de meetmomenten van het gevolgde proces voor de individuen zinvol te vergelijken zijn. Hiervoor is het noodzakelijk dat een ijkpunt aan te wijzen is in het gevolgde proces. Een voorbeeld is onderzoek naar het effect van een therapie, waarbij de start van de therapie als ijkpunt geldt. Ten tweede, het design waarbij de meetmomenten niet vergelijkbaar zijn tussen de individuen, en geen zinvol ijkpunt aangewezen kan worden in het proces dat gevolgd wordt. Een voorbeeld is onderzoek naar stemmingswisselingen in de loop der tijd bij een aantal individuen. In dit type design is het niet zinvol om naar gemeenschappelijke ontwikkelingsprocessen voor de verschillende individuen op zoek te gaan.

In sociaal wetenschappelijk onderzoek worden vaak multivariate data verzameld, waarbij de variabelen indicatoren zijn voor één of meerdere constructen. Een veel gebruikte analysemethode om de scores van meerdere individuen op meerdere variabelen efficiënt samen te vatten is principale componentenanalyse. De samenvatters van de variabelen zijn hopelijk, gewoonlijk na rotatie, interpreteerbaar in termen van de constructen die onderzocht worden.

Worden multivariate gegevens van verscheidene individuen longitudinaal verzameld, dan kunnen twee typen data onderscheiden worden, op basis van de eerder genoemde vergelijkbaarheid van meetmomenten. Bij 'longitudinale drie-weg data' zijn de meetmomenten zinvol vergelijkbaar tussen proefpersonen, terwijl bij 'multisubject multivariate tijdreeksen' dit niet het geval is. Dit proefschrift behandelt een aantal technieken voor componentenanalyse van zowel longitudinale drie-weg data als van multisubject multivariate tijdreeksen.

Longitudinale drie-weg data kunnen geanalyseerd worden met standaard componentenanalyse-modellen voor drie-weg gegevens, zoals CANDECOMP/PARAFAC (CP), Tucker3, Tucker2 en Tucker1, welke hiërarchisch geordend kunnen worden naar mate van strengheid. Afhankelijk van de data-structuur en de mate waarin samenvatting van de gegevens gewenst is, kan gekozen worden voor een meer of minder streng model. Een analyse met één van de twee strengste modellen, het CP- of Tucker3-model, van longitudinale drie-weg data levert drie componentenmatrices op, namelijk één voor de subjecten, één voor de variabelen, en één voor de tijdstippen. In de tijdscomponentenmatrix zijn de componentscores verzameld op de gemeten tijdstippen. Deze scores vatten het scoreverloop in de tijd voor de proefpersonen en de variabelen samen. Per component kunnen de scores op de verschillende tijdstippen beschouwd worden als evaluaties van een onderliggende functie, waarvan de exacte vorm meestal niet bekend is. In veel gevallen is het gerechtvaardigd om aan te nemen dat de functie een gladde vorm heeft. Soms kan echter aangenomen worden dat de scores een meer specifiek gedetermineerde functie volgen. Deze aannames kunnen gebruikt worden bij de analyse, door gladheidsrestricties, of een bepaalde functionele vorm, op te leggen aan de componentscores. Het voornaamste voordeel van het juiste gebruik van restricties is een reductie van de mate waarin het niet-structurele deel in de modelschattingen wordt opgenomen.

In Hoofdstuk 4 wordt voorgesteld om gladheidsrestricties op te leggen in de CP en Tucker3-modellen met behulp van 'B-splines'. B-splines kunnen op twee manieren aangewend worden om gladheid te bewerkstelligen in de CP- en Tucker3-modellen. Ten eerste kunnen de univariate reeksen met scores op opeenvolgende tijdstippen van de subjecten en de variabelen glad gemaakt worden, om daarna geanalyseerd te worden met een CP- of Tucker3-analyse. Ten tweede kunnen gladheidsrestricties opgelegd worden aan de opeenvolgende componentscores in de tijdscomponentenmatrix van het CP- of Tucker3-model. In een groot aantal gevallen hoeft de vraag welke van de twee benaderingen gebruikt dient te worden, niet gesteld te worden. Zoals in Hoofdstuk 4 bewezen wordt, leiden, onder bepaalde voorwaarden, deze twee methoden tot dezelfde modelschattingen. Additioneel aan gladheidsrestricties kunnen niet-negativiteits- en monotoniciteitsrestricties opgelegd worden met behulp van respectievelijk B-splines en de aan B-splines gerelateerde I-splines. In een simulatieonderzoek is de effectiviteit van het gebruik van gladheidsrestricties in de CP- en Tucker3-modellen onderzocht. Bij gesimuleerde drie-weg data met een gladde structuur bleek in het algemeen dat de parameters van de CP- en de Tucker3-modellen beter geschat worden als gladheidsrestricties worden opgelegd, dan wanneer een ongerestricteerd model gebruikt wordt. Een empirisch voorbeeld illustreert het gebruik van gladheidsrestricties in het Tucker3-model. Het Tucker3-model met gladheidsrestricties bleek eenvoudiger te interpreteren dan het ongerestricteerde Tucker3-model.

In Hoofdstuk 5 worden bepaalde functies aan de tijdscomponentscores in het Tucker3-model opgelegd. Het aldus verkregen model wordt het 'Structured Latent

Curve (SLC) Tucker3-model' genoemd. Het SLC Tucker3-model is geïnspireerd op de 'Structured latent curve' factormodellen voor de analyse van univariate longitudinale gegevens van meerdere proefpersonen, zoals voorgesteld door Browne en Du Toit (1991) en Browne (1993). In het SLC Tucker3-model is iedere tijdscomponent een, vooraf gespecificeerde, functie van een beperkt aantal parameters. Analooq aan de methode van Browne en Du Toit wordt een eerste orde Taylor-polynoom gebruikt van een bepaalde doelfunctie om de functies van de 'latente curves' te beschrijven met behulp van een beperkt aantal basiscurven. Als doelfunctie is gekozen voor de Gompertzfunctie, welke vooral nuttig is om groeigegevens te modelleren. De parameters van de Gompertzfunctie en de basisfuncties zijn goed interpreteerbaar. De eigenschappen van het SLC Tucker3-model, en een algoritme om het model aan data te fitten zijn besproken. Voor de empirische gegevens waarvan in Hoofdstuk 4 een Tucker3-model met gladheidsrestricties is besproken, is een SLC Tucker3-model geschat. De interpretatie van de parameters vertoonde grote overeenkomsten. Echter, het SLC Tucker3-model is spaarzamer, en verdient daarom hier de voorkeur.

De analyse van multisubject multivariate tijdreeksen vereist een andere aanpak dan de analyse van longitudinale drie-weg data. Omdat de meetmomenten niet vergelijkbaar zijn voor de verschillende individuen is het niet zinvol om de tijdreeksen voor verschillende individuen samen te vatten, zoals in de besproken modellen voor longitudinale drie-weg data wordt gedaan. In Hoofdstuk 6 en 7 worden modellen voor de analyse van multisubject multivariate tijdreeksen behandeld, waarin zowel intra-individuele variabiliteit als interindividuele variabiliteit wordt gemodelleerd. In Hoofdstuk 6 worden vier varianten van Simultane ComponentenAnalyse (SCA) besproken. In elk van de vier modellen worden de multivariate tijdreeksen ontbonden in een beperkt aantal tijdreeksen van componentscores, en een variabelen componentenmatrix, waarop de interpretatie van de componenten is gebaseerd. De vier modellen kunnen hiërarchisch geordend worden van zwak naar streng gerespecteerd. In het zwakste model wordt de meest variatie tussen individuen toegestaan, en in het strengste model de minste. Aandacht wordt besteed aan het schatten van de modelparameters, aan de keuze van het type model en het aantal componenten, en aan de rotatievrijheid van de vier modellen. Het gebruik van de SCA-modellen wordt geïllustreerd aan de hand van twee empirische voorbeelden.

In Hoofdstuk 7 worden de vier SCA-modellen uitgebreid, zodanig dat niet alleen gelijktijdige effecten, maar ook vertraagde effecten gemodelleerd kunnen worden. Deze vier 'Lagged SCA (LSCA)-modellen' zijn geïnspireerd op een variant van de klasse van dynamische factor-modellen voorgesteld door Molenaar (1986). Molenaar's dynamische factor-model wordt gewoonlijk gebruikt om multivariate tijdreeksen van één individu te modelleren, terwijl hier de multivariate tijdreeksen van meerdere individuen worden gemodelleerd. Algoritmen om de LSCA-modellen te fitten worden besproken, evenals de rotatievrijheid binnen de vier LSCA-modellen. Deze rotatievrijheid bleek, onder bepaalde assumpties, beperkt. In een klein

simulatieonderzoek zijn de schattingseigenschappen onderzocht van het algoritme voor het fitten van het minst gerespecteerde LSCA-model. De parameters van de gesimuleerde data bleken in het algemeen redelijk teruggeschat te worden, met uitzondering van de ‘moeilijke’ condities, bijvoorbeeld een hoog ruisniveau in de gesimuleerde gegevens. De analyse van een empirische data set leverde teleurstellende resultaten op, omdat de geschatte modelparameters nauwelijks interpreteerbaar bleken.

Hoofdstuk 8 sluit het proefschrift af met een samenvatting, een discussie van de gevonden resultaten, en mogelijke uitbreidingen van de gepresenteerde modellen. Ook wordt een mogelijke tekortkoming in de schattingsprocedures aangestipt.